PDF Download
3721981.pdf
28 January 2026
Total Citations: 1
Total Downloads: 978

Published: 23 May 2025
Online AM: 12 March 2025
Accepted: 23 February 2025
Revised: 15 January 2025
Received: 07 November 2024

Citation in BibTeX format

RESEARCH-ARTICLE

Towards Energy-efficient Audio-visual Classification via Multimodal Interactive Spiking Neural Network

XU LIU, Hefei University of Technology, Hefei, Anhui, China

NA XIA, Hefei University of Technology, Hefei, Anhui, China

JINXING ZHOU, Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, Abu Dhabi, United Arab Emirates

ZHANGBIN LI, Hefei University of Technology, Hefei, Anhui, China

DAN GUO, Hefei University of Technology, Hefei, Anhui, China

Open Access Support provided by:

Hefei University of Technology

Mohamed Bin Zayed University of Artificial Intelligence

ACM Transactions on Multimedia Computing, Communications and Applications, Volume 21, Issue 5 (May 2025)
https://doi.org/10.1145/3721981
EISSN: 1551-6865

Towards Energy-efficient Audio-visual Classification via Multimodal Interactive Spiking Neural Network

XU LIU and NA XIA, Hefei University of Technology, Hefei, China
JINXING ZHOU, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates
ZHANGBIN LI, Hefei University of Technology, Hefei, China
DAN GUO, Hefei University of Technology, Hefei, China and Hefei Comprehensive National Science Center, Hefei, China

The Audio-visual Classification (AVC) task aims to determine video categories by integrating audio and visual signals. Traditional methods for AVC leverage Artificial Neural Networks (ANNs) that operate on floating-point features, affording large parameter counts and consuming extensive energy. Recent research has shifted towards brain-inspired Spiking Neural Networks (SNNs), which transmit audiovisual information through sparser 0/1 spike features allowing for better energy efficiency. However, a byproduct of such sparsity is the increased difficulty in effectively encoding and utilizing these spike features. Moreover, the spike firing characteristics based on neuron membrane potential cause asynchronous spike activations due to the heterogeneous distributions of different modalities in the AVC task, resulting in cross-modal asynchronization. This issue is often overlooked by prior SNN models, resulting in lower classification accuracy compared to traditional ANN models. To address these challenges, we present a new Multimodal Interaction Spiking Network (MISNet), the first to successfully balance both accuracy and efficiency for the AVC task. As the core of MISNet, we propose a Multimodal Leaky Integrate-and-fire (MLIF) neuron, which coordinates and synchronizes the spike activations of audiovisual signals within a single neuron, distinguishing it from the prior paradigm of SNNs that relies on multiple separate processing neurons. As a result, our MISNet enables to generate audio and visual spiking features with effective cross-modal fusion. Additionally, we propose to add extra loss regularizations before fusing the obtained audio-visual features for final classification, thereby benefiting unimodal spiking learning for multimodal interaction. We evaluate our method on five audio-visual datasets, demonstrating advanced performance in both accuracy and energy consumption.

The work was supported in part by the National Key R&D Program of China under Grant No. 2024YFB3311602; National Natural Science Foundation of China (NSFC) under Grants 61971178, 62272144, and U20A20183; in part by Ordos Science and Technology Major “Open Bidding for Selecting the Best Candidates” Project under Grant JBGS-2023-002; in part by State Grid Co., Ltd. Headquarters Technology Project under Grant 5500-202140127; in part by the Major Project of Anhui Province under Grants 202423k09020001; in part by the Anhui Provincial Natural Science Foundation—Distinguished Young Scholars Project (2408085J040); and in part by the Fundamental Research Funds for the Central Universities under Grants JZ2024HGTG0309 and JZ2024AHST0337.

Authors’ Contact Information: Xu Liu, Hefei University of Technology, Hefei, China; e-mail: 1349087341liu@gmail.com; Na Xia (corresponding author), Hefei University of Technology, Hefei, China; e-mail: xiananawo@hfut.edu.cn; Jinxing Zhou (corresponding author), Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates; e-mail: jinxing.zhou@mbzuai.ac.ae; Zhangbin Li, Hefei University of Technology, Hefei, China; e-mail: lizhangbin.mail@gmail.com; Dan Guo (corresponding author), Hefei University of Technology, Hefei, China and Hefei Comprehensive National Science Center, Hefei, China; e-mail: guodan@hfut.edu.cn.

ACM 1551-6865/2025/5-ART144
https://doi.org/10.1145/3721981

CCS Concepts: • **Computing methodologies** → *Neural networks*; • **Information systems** → *Multimedia content creation*;

Additional Key Words and Phrases: Audio-Visual Classification, Spiking Neural Networks, Multimodal Interaction, Multimodal Leaky Integrate-and-Fire Neuron

ACM Reference format:
Xu Liu, Na Xia, Jinxing Zhou, Zhangbin Li, and Dan Guo. 2025. Towards Energy-efficient Audio-visual Classification via Multimodal Interactive Spiking Neural Network. *ACM Trans. Multimedia Comput. Commun. Appl.* 21, 5, Article 144 (May 2025), 24 pages.
https://doi.org/10.1145/3721981

1 Introduction

Artificial Neural Networks (ANNs) have achieved significant success in artificial intelligence applications such as computer vision [18], multi-modals [3, 9, 10, 40], by simulating the hierarchical structure of the visual cortex. However, their computational costs are extremely high [18]. A standard computer requires about 250 watts [31] of power to recognize objects and sounds, whereas the human brain consumes only about 20 watts when simultaneously performing complex tasks such as object and sound recognition, reasoning, control, and movement. Many real-world platforms, such as smartphones and Internet of Things devices, face constraints in terms of resources and battery life, which limits the implementation of deep neural networks [47]. To enable intelligence on these platforms, it is of great significance to explore how to utilize the inherently efficient computational paradigm of biological neural systems (i.e., brain-inspired computing) to achieve low-power implementations of neural networks.

Biologically inspired **Spiking Neural Networks (SNNs)** offer distinct advantages such as biological feasibility, event-driven sparsity, and binary activation [14], enabling them to achieve artificial intelligence as efficient as brain-inspired computing [27, 51]. The activated spike feature allows matrix multiplications to be converted into accumulation operations, further enhancing computational efficiency [25]. Unlike traditional ANN-based methods, SNN-based methods perform all forward computations entirely through binary 0/1 spike calculations, significantly reducing power consumption on hardware platforms such as GPUs. In addition, the SNN architecture is performed in an event-driven manner, which allows it to capture key information from dynamic event changes and perform computations only when events occur (i.e., the triggering of spikes), thus avoiding resource waste. Recent research has shown that SNNs have made significant progress in fields such as computer vision [41] and natural language processing [1]. Particularly, SNN-based architectures such as Spiking CNN [8] and Spiking Transformer [33, 46, 62] have demonstrated low energy consumption while maintaining excellent performance in vision tasks.

Although SNNs have shown their effectiveness in various unimodal tasks, their application in the multimodal domain remains relatively limited. In contrast to unimodal research tasks, the exploration of multimodal tasks, necessitating utilization and fusion of multiple modalities, can significantly enhance a model's ability to understand and process complex scenarios, making it an emerging and important research direction [12, 29]. **Audio-visual Classification (AVC)** is a typical yet straightforward task, as shown in Figure 1(a), which involves the fusion of visual (image) and auditory information for video classification, its practical applications include emotion recognition in human-computer interaction [30], autonomous driving [52]. These scenarios demand high multimodal fusion capabilities. Applying SNNs to these scenarios offers advantages such as low energy consumption and ease of deployment on embedded devices, making the exploration of SNN-based AVC methods highly significant [11, 55].

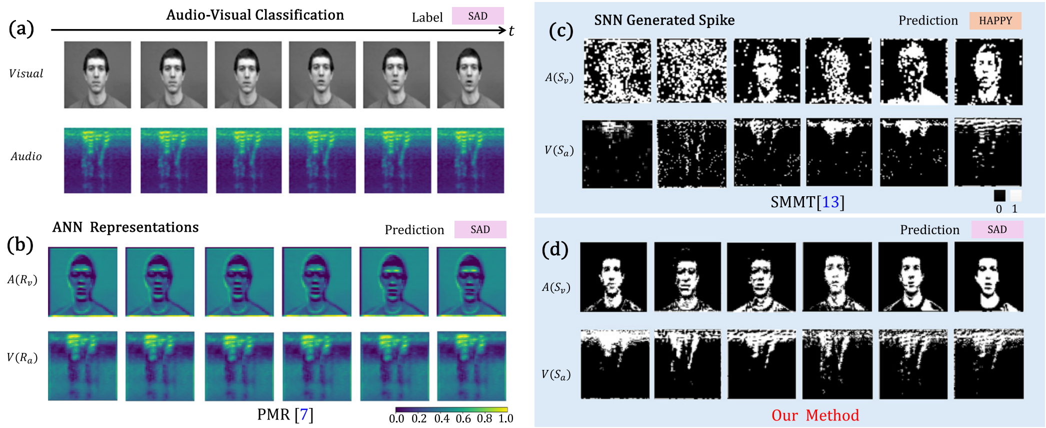
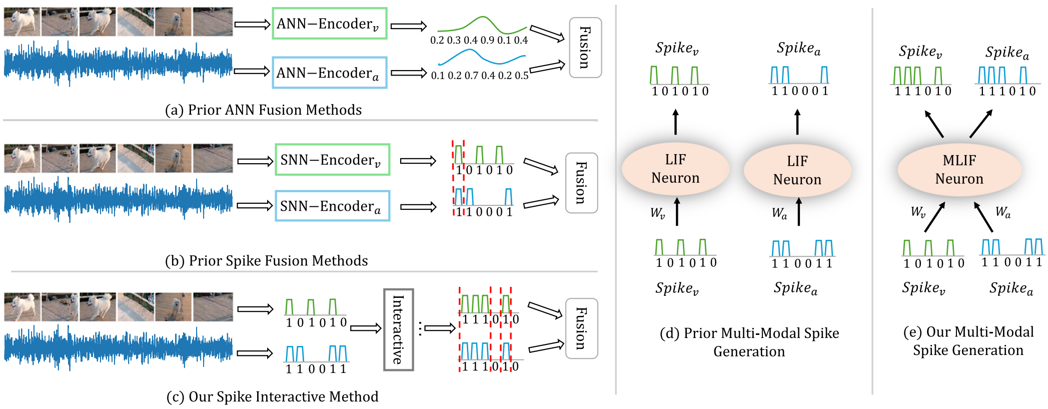


Fig. 1. (a) Illustration of the Audio-visual Classification (AVC) task which aims to determine the video event category by integrating audio and visual signals. (b) Prior ANN methods, for example, PMR [6], typically use continuous float-point values to embed audio-visual features, leading to extensive energy consumption. (c) The recent work, for example, SMMT [12], starts to employ SNN architecture whereas the audio and visual features are embedded by discrete binary spikes (0 or 1). This helps to reduce the energy consumption. However, we visualize the learned spike features and find they cannot satisfactorily capture key audio-visual information related to the video event across the timeline, causing incorrect predictions. (d) In contrast, the spike features learned by our proposed SNN-based method can effectively capture key audiovisual cues, such as facial expressions and relevant audio spectrum, striking a balance between accuracy and energy efficiency.

Traditional ANN-based methods use full-precision floating-point representations for all forward propagation, which requires substantial computational resources, as show in Figure 1(b). In contrast, SNN-based methods execute all forward propagation solely through 0/1 spikes, significantly reducing energy consumption in multimodal computations, as shown in Figure 1(c) and (d). Additionally, SNNs can leverage their excellent event-processing capabilities to optimize the handling of temporal information and dynamic changes in video data. Exploring the potential of SNNs in this task holds great value and could further extend to other related multimodal tasks [43–45, 50, 59].

However, directly applying SNNs to the multimodal AVC task is not easy. In traditional ANNs, as shown in Figure 2(a), visual and audio modalities are processed through two encoders, which output floating-point representations and map them to the same continuous domain for representation [29]; these rich floating-point representations effectively facilitate multimodal fusion [35]. In contrast, SNNs perform forward computation using 0/1 spikes, and their encoders extract simple spike features, as shown in Figure 2(b), which are fundamentally different from continuous feature representations. When integrating multimodal spike features using SNNs, two main challenges arise: (1) Compared to floating-point features, the spike features composed of 0s and 1s are sparser and coarser [62], making it more difficult to effectively encode pivotal cues from each modality. (2) Due to the heterogeneity between the audio and visual modalities, differences in the processing of these modalities lead to asynchronous spike emissions. This may lead to *unsynchronized spike activation* between the audio and visual modalities. As a result, the model may make predictions in advance according to the dominated activation from a single modality. For example, the prior SNN-based method SMMT [12] in the AVC task directly fuses the spike features of audio and visual modalities in the later stage but overlooks their asynchronous spike emissions across temporal steps. As shown in Figure 1(c), the spikes generated by the audio modality effectively capture some



(a) Prior ANN Fusion Methods

(b) Prior Spike Fusion Methods

(c) Our Spike Interactive Method

(d) Prior Multi-Modal Spike Generation

(e) Our Multi-Modal Spike Generation

Fig. 2. Examples of typical AVC task scenario, the dog appears only in a few frames, while the barking sound continues throughout, leading to a notable difference in the responses of the visual and audio modalities: (a) Prior ANN paradigms use separate encoders to represent video and audio s as floating-point features, mapping them into the same feature space to extract effective floating-point representations; (b) Prior SNN paradigms use two independent encoders to extract spike representations; however, due to differences between the visual and audio modalities, spike activations are asynchronous across modalities, affecting the final fusion outcome; (c) Our paradigm introduces a multi-round spike interaction mechanism in the AVC task. Through multiple interactions, it synchronizes the barking audio spikes and the dog's visual spikes to the same spike timestep, achieving consistent spike activations across modalities and thereby yielding improved results; (d) Prior methods in neuron-level only use two independent neurons to emit spikes, failing to capture the relationships between multimodal spike signals; (e) Our method at neuron level utilizes MLIF to achieve multimodal spike interaction, it can capture multimodal spike relationships.

key cues in the audio spectrum, while the visual modality emits a large number of irrelevant spikes. This limitation may stem from the constraints of the spike generation paradigm: From a macro level, this paradigm uses two independent spike encoders to extract spike features, resulting in severe temporal misalignment between audio and visual spike sequences, as show in Figure 2(b), which weakens the collaborative expression of bimodal features. For a micro level, this paradigm employs independent spiking neurons to emit spikes for audio and visual modalities separately, as show in Figure 2(d), causing the spike emissions of visual and audio modalities to be independent of each other, ultimately impacting the optimization of fusion performance.

To address these issues, we design a novel multimodal neuron called the **Multimodal Leaky Integrate-and-fire (MLIF)** neuron. This neuron, for the first time, integrates spike emissions from two modalities at the neuron level while maintaining the efficiency of binary computations. MLIF is primarily utilized to process multimodal data and achieve synchronized spike emissions. Unlike traditional **Leaky Integrate-and-fire (LIF)** neurons, the MLIF neuron integrates multimodal inputs within a single neuron. During the charging phase, MLIF configures multimodal membrane potentials within a single neuron, and during the discharging phase, these multimodal membrane potentials merge to emit spikes, as shown in Figure 2(e). Additionally, multimodal membrane potentials accumulate within the neuron over time T and influence each other. Through this approach, MLIF resolves the issue of asynchronous spike emissions caused by using two independent LIF neurons. Based on the MLIF neuron, we have constructed a new spiking multimodal paradigm and designed a novel **Multimodal Interactive Spiking Neural Network (MISNet)** for AVC tasks, marking the first introduction of multimodal interaction into SNNs. As shown in Figure 2(c), in MISNet, audio and visual modalities achieve spike interaction through our designed multimodal spike interaction units, which regulate spike emissions from the audio and visual modalities to achieve synchronized spike emissions. Although the final multimodal fusion occurs at the MISNet output, the multiple rounds of spike coordination in MISNet ensure that the fused spikes do not depend on any single modality, effectively mitigating the issue of imbalanced multimodal spike emissions. Furthermore, to more effectively learn the spike generation process for each modality, we calculate separate target losses for each modality during the training of MISNet and incorporate these losses as regularization terms. Through this joint optimization strategy, MISNet learns to generate spike features that are most relevant to the inputs. Figure 1(d) illustrates the spike generation process in MISNet, enabling the accurate generation of video and audio features through 0/1 spikes. Our main contributions can be summarized as follows:

- We propose the MISNet, a new multimodal SNN-based network for the AVC task. Our MISNet leverages multi-round interactions between audio and visual spike signals during forward computation for more effective multimodal learning.
- We design the MLIF neuron, a unique computational paradigm of multimodal SNN unit, that dynamically integrates and balances information from multiple modalities to emit unified spike sequences for final classification.
- Extensive experiments on five AVC datasets confirm the effectiveness of our proposed MISNet. Compared to existing SNN models and traditional ANN models, our architecture exhibited superior performance with greater parameter efficiency and lower energy consumption.

2 Related Work

2.1 AVC

AVC [12] aims to identify and classify events and activities in videos using both auditory and visual modalities, which is a fundamental research task in audio-visual learning [13, 21, 22, 26, 32, 53–55, 57–61]. AVC task is more complex than uni-modal audio classification or visual classification problems due to the heterogeneity between different modalities. Effective audio-visual fusion strategies are able to facilitate the AVC task. The early works simply adopt the recurrent neural networks for audio-visual fusion [5]. Later methods design various attention mechanisms to achieve the cross-modal fusion [29], whereas the multi-head attention proposed in Transformer [38] is widely used. Recently, Zhou et al. propose a positive sample propagation strategy [56] to identify and select only the highly relevant audio-visual pairs for fusion. Moreover, some works emphasize on balanced audio-visual fusion [39, 42, 48], as well as research on more complex tasks enabled by audio-visual multimodality, such as weakly supervised referring expression grounding [24] and change captioning [36, 49]. Although these methods are effective and beneficial for improving performance, they are mostly based on ANNs and are computationally intensive with high energy consumption. In contrast, our approach addresses the AVC task by designing a simple SNNs architecture striking a balance between accuracy and energy efficiency.

2.2 SNN

SNNs demonstrate substantial potential in the field of deep learning, mainly due to their neuron modeling and learning rules that draw inspiration from real biological mechanisms [16]. SNNs can effectively integrate techniques in typical ANNs, such as network architectures [62], training methodologies [63], gradient backpropagation [15, 20], and normalization [4, 17], significantly enhancing the performance in various tasks. Currently, the predominant training methods for SNNs include conversion methods from ANNs to SNNs [28] and direct training methods [63], with surrogate gradients being a popular approach for the latter. SNNs have shown comparable performance to ANNs in various unimodal tasks, such as image classification [46, 62] and natural language processing [1]. However, the complexity of multimodal tasks, especially the cross-modal

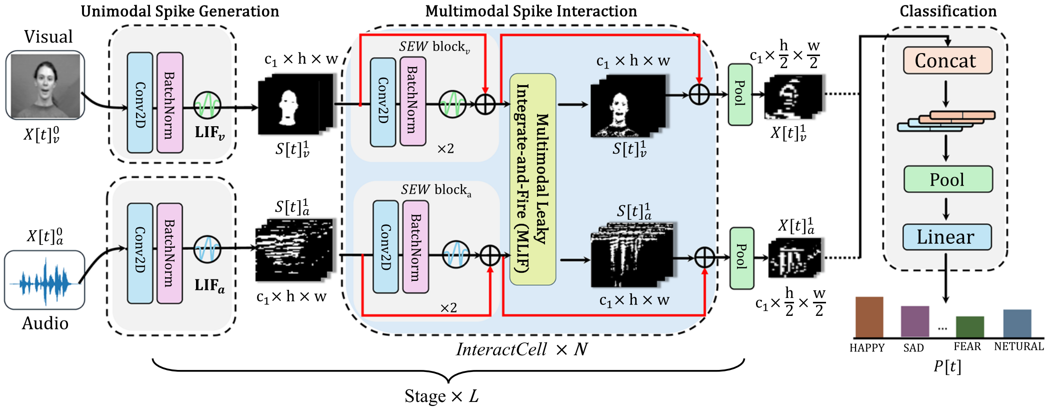


Fig. 3. Overview of our proposed MISNet for AVC. Our MISNet is primarily composed of L stages of encoders followed by a classification head. The encoder initially employs a *Unimodal Spike Generation* block, which separately extracts the spike representations for audio and visual modalities. Then, the *Multimodal Spike Interaction* block utilizes a core *MLIF* neuron to integrate audio and visual spikes, achieving synchronized multimodal spike learning.

heterogeneity in the studied AVC task, makes the SNNs less competitive than ANNs. For example, SMMT [12] attempts to address the AVC task using a SNN. Although the utilization of SNN helps to save energy, this method still suffers from large parameter overhead, and the performance/accuracy is far from the state-of-the-art ANN method. In summary, prior works simply introduce binary spike activations, which fails to consider the unique characteristics of the audio and visual modalities and the cross-modal spike activation imbalance issue.

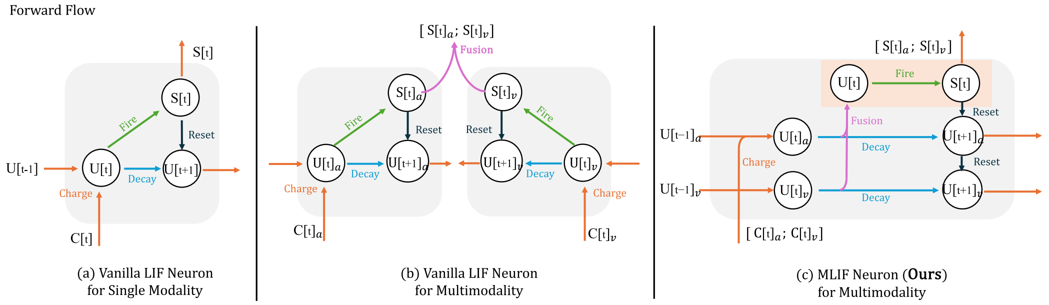
3 Method

As illustrated in Figure 3, our proposed MISNet primarily consists of L stages of encoders and a classification head. Each stage is composed of two core components: (1) *Unimodal Spike Generation*, which describes how high-dimensional mappings are applied to the audio and visual inputs, encoding each modality into 0/1 spike features; (2) *Multimodal Spike Interaction*, which utilizes our proposed MLIF neuron to enable multimodal interactive learning and generate synchronized spike features. The Pooling layer is used to downsample the audio/visual spike features to reduce dimensionality and computational load before sending them to the next stage. Finally, after the multimodal interaction through the encoders, the generated audio and visual spike features are concatenated and sent to a linear layer to predict the video category.

3.1 Preliminary

The fundamental unit of SNNs is the spike neuron, which accumulates membrane potential from the input postsynaptic current and compares it to a pre-set threshold to generate spikes. The LIF [41] is a typical neuron for spike generation. We provide an illustration of LIF in Figure 4(a). The whole process can be formulated as a differential equation as follows:

$$\tau \frac{du(t)}{dt} = -u(t) + RI(t). \tag{1}$$



(a) Vanilla LIF Neuron for Single Modality

(b) Vanilla LIF Neuron for Multimodality

(c) MLIF Neuron (**Ours**) for Multimodality

Fig. 4. Illustration of the differences between MLIF and LIF. (a) Vanilla LIF neuron is used to process a single modality, with a neuron handling the signals of that modality; (b) Vanilla LIF neuron is used for multimodality, employing two independent neurons to process the signals of each modality separately; (c) Our proposed MLIF neuron is designed for multimodality, capable of synchronous processing multiple modality inputs within a single neuron.

Its iterative form can be represented by the following equations:

$$\begin{cases} U\,[t] = (1 - \frac{1}{\tau})U\,[t-1] + C\,[t]\,, & \textbf{Charge} \\ S[t] = \Psi\,(U\,[t] - V_{\mathrm{t}}\,) = \begin{cases} 1, & \text{if } U\,[t] \geq V_{\mathrm{t}} \\ 0, & \text{otherwise} \end{cases} & , \textbf{Fire} \\ U\,[t+1] = U\,[t] - V_t S\,[t]\,, & \textbf{Reset}. \end{cases} \tag{2}$$

Specifically, the operations of the LIF neuron involve three key steps: (1) *Charge*: The neuron integrates the input postsynaptic current $C[t]$ and membrane time constant τ, simulating the process of a capacitor accumulating voltage over time, which leads to the accumulation of membrane potential *U[t]*; (2) *Fire*: When $U[t]$ reaches or exceeds the given threshold V_{th} through the Heaviside function $\Psi(\cdot)$, the neuron fires, emitting a spike; otherwise, no spike is emitted; (3) *Reset*: After a spike is emitted, $U[t]$ undergoes a reset and decay, simulating the recovery process of biological neurons.

3.2 Unimodal Spike Generation

Given the visual frames and audio sequence, we first generate the unimodal spike representations for each modality. This is achieved through the aforementioned LIF neuron and some learnable layers including Conv2D and BatchNorm ($\mathcal{BN}$). Specifically, for the timestep $t \in \{1, 2, \ldots, T\}$ in the lth stage encoder $l \in \{1, 2, \ldots, L\}$, the model receives the output from the prior stage and generates spike features for each modality. Let us denote the outputs of visual and audio modalities from the prior $l-1$ stage as $X[t]_v^{l-1} \in \mathbb{R}^{c_{l-1} \times h_{l-1} \times w_{l-1}}$ and $X[t]_a^{l-1} \in \mathbb{R}^{c_{l-1} \times h_{l-1} \times w_{l-1}}$, respectively. Here, c_{l-1} represents the channel dimension of features, h_{l-1} and w_{l-1} denote the height and width of the feature map, respectively. At the first stage ($l = 1$), $X[t]_v^0$ and $X[t]_a^0$ correspond to the initial visual and audio inputs, respectively. Then, the spike generation process can be described as follows:

$$\begin{aligned} C[t]_m^l &= \mathcal{BN}(\mathrm{Conv2D}_m^l(X[t]_m^{l-1})), \\ S[t]_m^l &= \mathrm{LIF}_m(C[t]_m^l), \end{aligned} \tag{3}$$

where $m \in \{a, v\}$ denotes the audio/visual modality and $C[t]$ represents the postsynaptic current, simulated through a learnable Conv2D layer to mimic synaptic firing. $S[t]_m^l \in \mathbb{R}^{c_l \times h_l \times w_l}$ is the generated spike binary features.

3.3 Multimodal Spike Interaction

The unimodal spike representations $S[t]_m^l$ are obtained independently. We further consider enhancing the multimodal interactions between audio and visual spike features to utilize the unsynchronized spike activations across modalities.

As shown in Figure 3, the unimodal spike features at the lth stage $S_m^l[t]$ are further processed by N multimodal spike interactions. We name the multimodal interaction at nth round of the lth stage as *InteractCell*^{l,n}, where $n \in \{1, 2, ..., N\}$ and $l \in \{1, 2, ..., L\}$. Next, we provide details of each *InteractCell*^{l,n}. Notably, we abbreviate the superscript n in the following equations for ease of expression. Specifically, the unimodal spike features $S[t]_m^l$ are first passed through two *SEW* [8] blocks, each comprising "Conv2D-$\mathcal{BN}$-LIF" layers with a residual connection. This process can be described as

$$\begin{aligned} O[t]_m^l &= S[t]_m^l + SEW(S_m^l[t]), \\ SEW(S_m^l[t]) &= \text{LIF}(\mathcal{BN}(\text{Conv2D}(\text{LIF}(\mathcal{BN}(\text{Conv2D}(S_m^l[t])))))), \end{aligned} \tag{4}$$

where $O[t]_m^l \in \mathbb{R}^{c_l \times h \times w}$ is the output from the *SEW* blocks.

Next, $O[t]_m^l (m \in \{a, v\})$ is fed into our proposed *MLIF neuron* to update the spike representations by considering cross-modal relations, written as

$$\begin{aligned} C[t]_m^l &= O[t]_m^l \cdot W_m^l, \\ S[t]_v^l, S[t]_a^l &= \text{MLIF}(C[t]_v^l, C[t]_a^l), \end{aligned} \tag{5}$$

where $W_m^l \in \mathbb{R}^{h \times w}$ is a learnable weight matrix used to transform $O[t]_m^l$ into the input postsynaptic current $C[t]_m^l$ required for the multimodal interaction process. $S[t]_v^l$ and $S[t]_a^l$ are the updated visual and audio spike features, respectively. Next, we elaborate on the core MLIF neuron.

MLIF Neuron. As shown in Figure 4(c), unlike the traditional approach of using two LIF neurons to process the multimodalities independently, like Figure 4(b), our proposed MLIF neuron processes multimodal spike voltages and emits 0/1 spikes by jointly processing the postsynaptic current $C[t]_m^l$ from multiple modalities, rather than firing independently. The operations of the MLIF neuron are based on the multimodal postsynaptic current and corresponding membrane potentials, consisting of *four* core steps: *Multimodal Charge*, *Fusion*, *Fire*, and *Reset*. In the following, we explain each step in detail.

(1) *Multimodal Charge.* Since the changes of membrane potential are influenced by multiple modalities, their dynamic fusion must be considered during the charging process. For the tth timestep, given the input postsynaptic current from each modality $C\,[t]_m$ and the prior membrane potential at $t-1$ timestep $U\,[t-1]_m$, the membrane potential at tth timestep $U\,[t]_m^l$ can be updated/charged as follows:

$$U\,[t]_m^l = \left(1 - \frac{1}{\tau}\right) U\,[t-1]_m^l + C\,[t]_m^l, \tag{6}$$

where τ is a membrane time constant, empirically set to 0.2. $m \in \{a, v\}$ denotes the audio or visual modality.

(2) *Multimodal Fusion.* The above charging process still responds independently within each modality. To achieve a genuine integration of membrane potentials from audio and visual modalities simultaneously, we fuse $U\,[t]_v^l$ and $U\,[t]_a^l$ through an automatic mechanism, formulated as

$$U\,[t]^l = \left[\alpha_v^l U\,[t]_v^l ; \alpha_a^l U\,[t]_a^l\right], \tag{7}$$

where α_v^l and α_a^l are two learnable parameters, and $[\cdot;\cdot]$ represents the concatenate (Concat) operation. $U\,[t]^l \in \mathbb{R}^{2c_l \times h \times w}$ is the membrane potential after multimodal fusion.

(3) *Multimodal Fire.* After obtaining $U\,[t]^l$, we can generate the spike $S\,[t]^l$ by comparing $U\,[t]^l$ with a pre-set threshold V_t. If $U\,[t]^l$ is greater than V_t, the membrane potential will fire to generate a positive spike (value +1); otherwise, the spike value will remain zero, written as

$$S\,[t]^l = \Psi(U\,[t]^l - V_t) = \begin{cases} 1, & \text{if } U[t]^l \geq V_t \\ 0, & \text{otherwise} \end{cases}. \tag{8}$$

Then, the obtained spike $S\,[t]^l \in \mathbb{R}^{2c_l \times h \times w}$ is split into two components, corresponding to the updated audio spike $S\,[t]_a^l \in \mathbb{R}^{c_l \times h \times w}$ and visual spike $S\,[t]_v^l \in \mathbb{R}^{c_l \times h \times w}$, which can be further used as the inputs of the next MLIF neuron, described as

$$\left[S\,[t]_v^l\,; S\,[t]_a^l\right] = spilt(S\,[t]^l). \tag{9}$$

(4) *Multimodal Reset.* After the firing, it is necessary to reinitialize the neuron's state, simulating the periodic discharge and energy recovery of biological neurons, which is expected to recharge and re-spike at the appropriate time in the future. Under the multimodal condition, the reset process needs to be applied to each membrane potential $U[t]_m^l$, allowing for adjustments of contributions from each modality. The reset process can be operated following:

$$U\,[t+1]_m^l = U\,[t]_m^l - V_t S\,[t]_m^l, \tag{10}$$

where the initial membrane potential $U\,[t]_m^l$ is decayed by the multiplication of V_t with activated spike $S\,[t]_m^l$.

With the above principles, the processes of our MLIF neuron can be summarized as

$$\begin{cases} U\,[t]_m^l = (1 - \frac{1}{\tau})U\,[t-1]_m^l + C\,[t]_m^l, & \textbf{Multimodal Charge} \\ U\,[t]^l = \left[\alpha_v^l U\,[t]_v^l\,; \alpha_a^l U\,[t]_a^l\right], & \textbf{Multimodal Fusion} \\ S\,[t]^l = \Psi(U\,[t]^l - V_t) = \begin{cases} 1, & \text{if } U[t]^l \geq V_\text{t} \\ 0, & \text{otherwise} \end{cases}, & \textbf{Multimodal Fire} \\ \left[S\,[t]_v^l\,; S\,[t]_a^l\right] = spilt(S\,[t]^l), & \\ U\,[t+1]_m^l = U\,[t]_m^l - V_t S\,[t]_m^l, & \textbf{Multimodal Reset} \end{cases} \tag{11}$$

where $U\,[t-1]_m^l$ and $C\,[t]_m^l$ are the inputs of MLIF neuron, and $S[t]_m$ is the generated spikes.

3.4 Model Training and Inference

As shown in Figure 3, the entire MISNet consists of L stages. Given the inputs at lth stage, i.e., $X[t]_a^{l-1} \in \mathbb{R}^{C_{l-1} \times h_{l-1} \times w_{l-1}}$ and $X[t]_v^{l-1} \in \mathbb{R}^{C_{l-1} \times h_{l-1} \times w_{l-1}}$, the computation process for this stage can be expressed as

$$X[t]_a^l, X[t]_v^l = \text{Stage}^l(X[t]_a^{l-1}, X[t]_v^{l-1} | \Theta_l), \tag{12}$$

where $l = \{1, 2, ..., L\}$. Θ_l represents the learnable parameters for unimodal spike generation and multimodal spike interaction at lth stage. $X[t]_m^l \in \mathbb{R}^{c_l \times h_l \times w_l} (m \in \{a, v\})$ are the updated audiovisual spike features. Let us denote H and W as the height, and width of the initial audio and visual feature maps, we have $h_l = H/2^l$ and $w_l = W/2^l$. Channel dimension c_l is different in each stage.

The audio and visual spike features at the last Lth stage are concatenated to predict category probability $P[t]$ at tth timestep, computed as

$$P[t] = \text{Linear}(\text{Pool}(\text{Concat}(X[t]_a^L, X[t]_v^L))). \tag{13}$$

Here, $P[t] \in \mathbb{R}^{1\times C}$ (C is the total number of classes). In Equation (13), we use the Concat operation as the default fusion strategy and we will perform an ablation study on other fusion methods in Section 4.5.

Notably, as a SNN-based method, our proposed MISNet needs to successively process audio signals and visual frames through multiple timesteps. The final prediction result can be calculated by accumulating the outputs over T steps and dividing by T, formulated as, $p = \frac{\sum_{t=1}^{T} P[t]}{T} \in \mathbb{R}^{1\times C}$. Given the ground truth $Y \in \mathbb{R}^{1\times C}$, we can compute the MSE between p and Y as the basic loss, written as

$$\mathcal{L}_{mm} = MSE(p, Y). \tag{14}$$

$\mathcal{L}_{mm}$ regularizes the prediction obtained from multimodal fused spike features. Moreover, we propose to add constraints on unimodal spike features since the video event usually appears in both audio and visual modalities for the studied AVC task. To this end, we send the unimodal spike features $(X[t]_a^L, X[t]_v^L)$ at each timestep into independent linear layers to generate audio and visual event probability, denoted as $O[t]_a$ and $O[t]_v$, respectively. Then, the probability across T time steps can be computed by: $o_a = \frac{\sum_{t=1}^{T} O[t]_a}{T}$ and $o_v = \frac{\sum_{t=1}^{T} O[t]_v}{T}$. In this way, we can compute the unimodal loss as follows:

$$\mathcal{L}_a = \text{MSE}(o_a, Y), \quad \mathcal{L}_v = \text{MSE}(o_v, Y). \tag{15}$$

These loss items are beneficial for maximizing the retention of unimodal spiking during multimodal optimization.

The total objective for model optimization is calculated by summarizing the above three losses:

$$\mathcal{L} = \mathcal{L}_{mm} + \beta_a \mathcal{L}_a + \beta_v \mathcal{L}_v. \tag{16}$$

Notably, in the inference phase, only the predictions obtained from the multimodal spikes (i.e., p) are used. The final prediction can be determined by identifying the category having the highest probability value. We provide a comprehensive overview of our MISNet in Algorithm 1.

4 Experiments

4.1 Datasets

To demonstrate the efficacy and efficiency of our MISNet, we conduct experiments on two audio-image datasets, following prior works [12, 23]: CIFAR10-AV [12], Urbansound8K-AV (Ub8k-AV) [12], and the audio-visual sensor datasets MNIST-DVS and N-TIDIGITS (M&N) [23].

Next, we additionally consider more complex audio-video datasets, such as CREMA-D [2] and AVE [34], to further evaluate the applicability of MISNet. The CREMA-D dataset includes 6 emotional categories with a total of 7,442 clips, randomly divided into 6,698 training samples and 744 testing samples. The AVE dataset contains 28 event categories and 4,143 10-second videos.

Audio-visual Data Preprocessing. For the audio modality, we convert it into a Mel spectrogram. For the visual modality, we preprocess it following established standards [12, 46, 62]. For both the audio and video modalities, we set data to different input sizes across datasets: 32 × 32 for CIFAR10-AV, 96 × 96 for Urban8K-AV, and 128 × 128 for both CREMA-D and AVE. In CREMA-D and AVE, we extract 10 frames from each video clip and randomly select 4 frames to construct the training dataset.

4.2 Experimental Setup

MISNet Variants. We constructed four versions of the model—XS, S, L, and XL—which differ in channel dimensions and the number of stacked layers (see Table 1 for details). MISNet-L/XL can

```
Algorithm 1: MISNet for AVC
Input: Training data $\mathcal{D} = \{X_a, X_v\}$, Label $Y$ in mini-batch
Output: Audio-Visual Classification Loss $\mathcal{L}$
Initialize the $X_a^0, X_v^0 = X_a, X_v$, $P[1], \cdots, P[T] = 0$, $\{O[1]_a, O[1]_v\}, \cdots, \{O[T]_a, O[T]_v\} = 0$
 /* T × timestep */
for $t$ in $\{1, \cdots, T\}$ do
     /* L × Stage */
    for  $l$ in $\{1, \cdots, L\}$ do
        for $m$ in $\{a, v\}$ do    ▷ Unimodal Spike Generation
            Generate unimodal spikes $S[t]_m^l$ from $X[t]_m^l$ by Eq.3
        end for
        for $n$ in $\{1, \cdots, N\}$ do    ▷ Multimodal Spike Interaction
            for $m$ in $\{a, v\}$ do
                SEW block residual spike mapping for $O[t]_m^l$ by Eq.4
                Update postsynaptic current $C[t]_m^l$ from spike $O[t]_m^l$ by Eq.5(1)
            end for
            for $m$ in $\{a, v\}$ do    ▷ Multimodal Charge
                Update $U[t]_m^l$ from $C[t]_m^l$ by Eq.6
            end for
            Fusion $U[t]^l$ using $U[t]_a^l$ and $U[t]_v^l$ by Eq.7    ▷ Multimodal Fusion
            Fire spikes $S[t]^l$ from $U[t]^l$ by Eq.8    ▷ Multimodal Fire
            Split $S[t]^l$ into $S[t]_v^l$ and $S[t]_a^l$ by Eq.9
            for $m$ in $\{a, v\}$ do    ▷ Multimodal Reset
                Reset $U[t]_m^l$ by Eq.10
            end for
            for $m$ in $\{a, v\}$ do
                Downsample $S[t]_m^l$ to output $X[t]_m^l$ for the next Stage
            end for
        end for
    end for
    Current prediction $P[t]$ from $X[t]_a^L$ and $X[t]_v^L$ by Eq.13
    for $m$ in $\{a, v\}$ do
        Predict the current unimodal $O[t]_m$ from $X[t]_m^L$ using independent linear layers.
    end for
end for
$p = \frac{\sum_{t=1}^{T} P[t]}{T}$, $o_a = \frac{\sum_{t=1}^{T} O_a[t]}{T}$, $o_v = \frac{\sum_{t=1}^{T} O_v[t]}{T}$    ▷ Multimodal and Unimodal probability
Compute $\mathcal{L}_{mm}$, $\mathcal{L}_a$, $\mathcal{L}_v$ from $p$, $o_v$, $o_a$ by Eq. 14 and 15    ▷ Multimodal and Unimodal loss compute
Compute the total loss $\mathcal{L}$ by combining $\mathcal{L}_{mm}$, $\mathcal{L}_a$, and $\mathcal{L}_v$ using Eq.16
```

achieve better performance and is suited for more complex multimodal audio-visual tasks, offering improved handling of high-dimensional, high-complexity input data. However, its parameter count and computational complexity also increase accordingly. On the other hand, XS/S versions are optimized for simpler tasks, with reduced parameters and lower computational demands.

Implementation Details. In all comparison experiments, we set $\alpha_a^l = \alpha_v^l = 0.5$ according to Equation (9), and these parameters are initialized to 0.5 and subsequently updated and optimized through gradient descent during training. All models are trained using a mini-batch size of *32* with

Table 1. Detailed Architecture of MISNet Variants, Illustrating the Number of Stages L, the Channel Mappings for Each Stage ($\{c_1, c_2, \ldots, c_l\}$), and the Number of *InteractCell* per Stage

Model	L	$\{c_1, c_2, \ldots, c_L\}$	*InteractCell* $\times N$
MISNet-XS	3	$\{64, 128, 256\}$	$\{\times 2, \times 2, \times 1\}$
MISNet-S	3	$\{64, 128, 256\}$	$\{\times 2, \times 2, \times 2\}$
MISNet-L	4	$\{64, 128, 256, 512\}$	$\{\times 2, \times 2, \times 2, \times 1\}$
MISNet-XL	4	$\{64, 128, 256, 512\}$	$\{\times 2, \times 2, \times 2, \times 2\}$

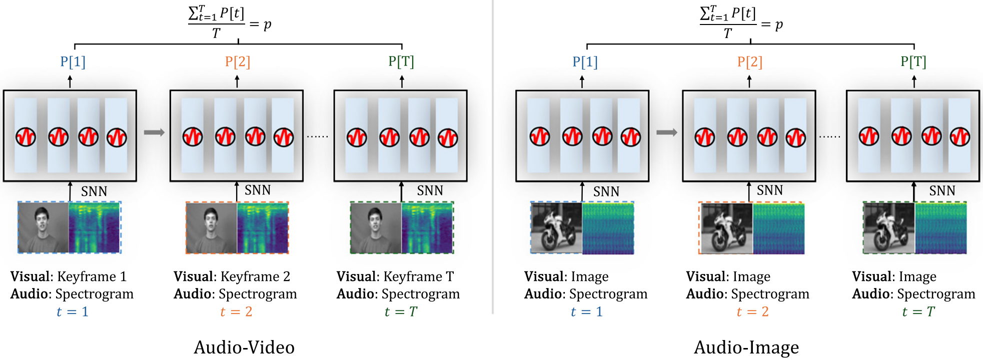


Fig. 5. Illustration of how the SNN processes different data datasets in different timesteps T, *Left*: Audio-video classification datasets (audio-visual sensor datasets), including CREMA-D [2], AVE [34], and MNIST-DVS, and N-TIDIGITS [23], where each keyframe and spectrogram of the input is processed as an input for a timestep t; *Right*: Audio-image classification datasets, including CIFAR10-AV [12] and Urbansound8k-AV [12], follow prior work [12, 23], where the same image and audio are input across different timesteps t.

the SGD optimizer, employing a momentum of 0.9 and a weight decay of 1×10^{-4}. The learning rate starts at 1×10^{-3} and gradually decays to 1×10^{-5}.

Surrogate Gradient. Due to the non-differentiability of the *Fire* operation in Equation (2) and the *Multimodal Fire* operation in Equation (11), MISNet follows previous surrogate gradient training methods during backpropagation, with *Sigmoid* selected as the surrogate function.

Platform. All experiments are conducted on a single NVIDIA A100 for both training and evaluation. We implement our model using PyTorch, with the SNN part utilizing the SpikingJelly [7] framework.

Timesteps. The SNN is set by default to emit spikes over a total of T timesteps. For AVC tasks, the input at each timestep $t \in T$ is different. Therefore, in the audio-video dataset, each timestep corresponds to a keyframe. Figure 5 illustrates the variation in audio-visual input across different time steps.

4.3 Performance Comparison with Prior Methods

Visual-only Dataset. As shown in Table 2, we also compared MISNet with recent unimodal SNN architectures, including Spikeformer-4-384 [62], DSpikeformer-4-384 [46], and Resformer-Ti [33], which are primarily used for visual classification. The comparison results indicate that MISNet not only requires fewer parameters than these models but also significantly outperforms them. For

Table 2. Performance Comparison of Existing SNN-base Methods on Visual-only and Audio-visual Classification

Modality	Architecture	Venue	Param (M) ↓	Timestep (T)	Acc (%) ↑	
					CIFAR10	Ub8k
Visual-only	ResNet18 [20]	*NIPS 2021*	11.69	2	93.13	-
	Spikeformer-4-384 [62]	*ICLR 2023*	9.32	4	95.19	-
	DSpikeformer-4-384 [46]	*NIPS 2023*	9.32	4	95.60	-
	Resformer-Ti [33]	*NIPS 2023*	10.79	4	96.24	-
	MISNet-XS (ours)	-	**4.06**	4	**96.32**	**89.04**
	MISNet-S (ours)		**7.33**	4	**96.61**	**90.11**
	MISNet-L (ours)		19.63	4	**96.82**	**91.05**
	MISNet-XL (ours)		31.00	4	**97.06**	**91.92**
Modality	**Architecture**	**Venue**	**Param (M) ↓**	**Timestep (T)**	**Acc (%) ↑**	
					CIFAR10-AV	**Ub8k-AV**
Audio-image	SMMT [12]	*TCDS 2023*	9.99	1	93.53	94.30
				4	97.01	96.85
	MISNet-XS (ours)	-	**4.06**	1	**99.03**	**96.91**
	MISNet-XS (ours)		**4.06**	4	**99.53**	**97.71**
	MISNet-S (ours)		**7.33**	1	**98.62**	**97.94**
	MISNet-S (ours)		**7.33**	4	**99.15**	**98.09**
	MISNet-L (ours)		19.63	1	**99.18**	**98.52**
	MISNet-XL (ours)		31.00	1	**99.26**	**98.96**

Our method achieved the best results on all datasets. The bold indicates the best performance.

instance, compared to Resformer-Ti, which achieves an accuracy of 96.24% at $T = 4$, our MISNet-XS reaches an accuracy of 96.32%. This further emphasizes that using MISNet can achieve better results compared to unimodal modals without requiring a large number of parameters.

Audio-image Datasets. Table 2 presents a comparison between MISNet and previous SNN-based models on Audio-image datasets, showing that MISNet outperforms the prior SMMT on both CIFAR10-AV and Urbansound8k-AV (Ub8k-AV) datasets. Specifically, MISNet-XS achieves a 5.5% improvement in accuracy over SMMT on CIFAR10-AV and a 2.60% improvement on Ub8k-AV, using fewer time steps T. Additionally, another configuration, MISNet-S, although slightly increased in parameter count compared to MISNet-XS, still requires fewer parameters than SMMT while achieving significantly better performance.

Audio-visual Sensor Dataset. The experimental results on the Audio-visual Sensor M&N dataset shown in Table 3 indicate that both MISNet-XS and MISNet-S reach an accuracy of 99.98%, surpassing prior methods SMMT [12] and EMSNN [23].

Intuitively, while SMMT performs well on CIFAR10-AV, it performs less effectively on the more complex Ub8k-AV dataset. This is mainly due to SMMT's use of two spiking encoders to extract spiking features, which are then fused via an attention mechanism. The complexity of images in Ub8k-AV makes it challenging for the visual modality to process these inputs effectively. In contrast, our MISNet demonstrates stable performance across both datasets.

Audio-video Datasets. As shown in Table 4, to further validate the effectiveness of MISNet, especially on complex audio-visual datasets, we conducted additional experiments on audio-video datasets (CREMA-D and AVE). The PMR [6] and QMF [50] methods are ANN-based approaches.

Table 3. Performance Comparison on the M&N Dataset, Where MISNet Achieves a *99.80%* in MISNet-S/L/XL Configurations

NNs type	Structure	Method	Acc (%) ↑
SNN	CNN-RNN	EMSNN [23]	99.10
	Multi-model Transformer	SMMT [12]	99.82
	CNN	**MISNet-XS (ours)**	**99.88**
	CNN	**MISNet-S (ours)**	**99.98**
	CNN	**MISNet-L (ours)**	**99.98**
	CNN	**MISNet-XL (ours)**	**99.98**

The bold indicates the best performance.

Table 4. Performance Comparison of State-of-the-art ANN and SNN Methods on AVC, Demonstrating That MISNet Significantly Outperform ANNs and Other SNN Methods

NNs type	Method	Venue	Param (M)↓	Acc (%) ↑	
				CREMA-D	AVE
ANN	Concat [6]	*CVPR 2023*	22.36	51.70	65.40
	Concat+Drop (audio) [6]	*CVPR 2023*		54.40	66.40
	Concat+Drop (visual) [6]	*CVPR 2023*		53.30	66.20
	PMR [6]	*CVPR 2023*	22.36	61.10	67.10
	QMF [50]	*ICML 2023*		63.71	-
SNN	SMMT [12]	*TCDS 2023*	9.99	62.50	51.78
	MISNet-XS (ours)	-	**4.06**	**66.45**	58.44
	MISNet-S (ours)		**7.33**	**68.72**	65.31
	MISNet-L (ours)		**19.63**	**75.22**	**67.24**
	MISNet-XL (ours)		**31.00**	**77.14**	**68.04**

The bold indicates the best performance.

As anticipated, MISNet achieved optimal results on both CREMA-D and AVE. For example, on the CREMA-D dataset, MISNet-S reached an accuracy of 68.72% using only 7.33M parameters, outperforming PMR [6] and QMF [50] by 7.61% and 5.08%, respectively. Additionally, we further explored the results of MISNet-L/XL. MISNet-L achieved an accuracy of 75.22% on the CREMA-D dataset and 67.24% on the AVE dataset. MISNet-XL achieved even higher performance, reaching 77.14% on CREMA-D and 68.04% on AVE. The results indicate that MISNet significantly outperforms ANN-based methods and the prior SMMT [12] in handling complex, multi-faceted AVC tasks, especially on datasets containing sequential frames. This superior performance is largely attributed to MISNet's unique multimodal spike interaction mechanism and its remarkable capability in temporal sequence capture, enhancing its effectiveness in processing dynamic audio-visual data.

4.4 Energy Estimation Comparison with Prior Methods

Table 5 highlights the significant advantages of MISNet in terms of energy efficiency and accuracy. As a high-efficiency neural network architecture, MISNet primarily transmits feature maps through sparse discrete spikes during forward propagation, effectively reducing energy consumption. We

Table 5. Energy Consumption Comparison on the Ub8k-AV and CREMA-D Datasets Shows That MISNet Achieves High Accuracy with Excellent Energy Efficiency

	Method	Ub8k-AV		CREMA-D	
		Acc (%) ↑	Energy (mJ) ↓	Acc (%) ↑	Energy (mJ) ↓
ANN	PMR [6]	-	-	61.10	266.60
	QMF [50]	-	-	63.71	266.60
SNN	SMMT [12]	96.85	3.99	62.50	5.78
	MISNet-XS (ours)	**97.71**	**0.97**	**66.58**	**1.32**
	MISNet-S (ours)	**98.09**	**1.76**	**68.72**	**2.41**
	MISNet-L (ours)	**98.52**	**2.55**	**75.22**	**4.58**
	MISNet-XL (ours)	**98.96**	4.23	**77.14**	6.88

The bold indicates the best performance.

Table 6. Ablation Study on the Final Classification Fusion Methods in the Ub8K-AV Dataset (Acc %), as Described in Equation (13), Where MISNet Consistently Outperforms SEW-ResNet (Two Independent Encoders for Spike Emission) across All Fusion Methods

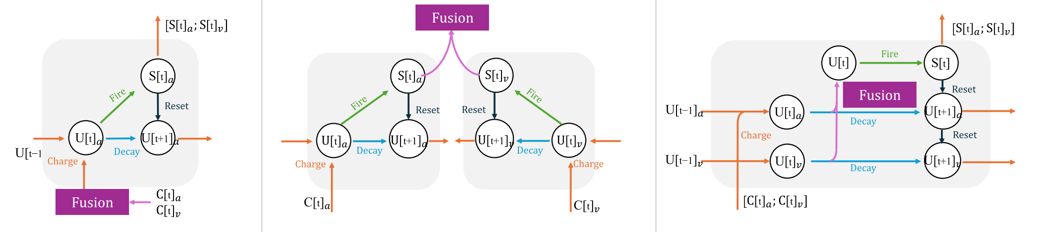
Architecture	Fusion				
	Sum	Hadamard	MLP	Attention	Concat
MISNet-S	**98.34**	**99.18**	**99.18**	**99.58**	**98.09**
SEW-ResNet18-EF	90.23	88.63	88.21	87.65	89.69
SEW-ResNet18-LF	85.72	85.21	89.21	90.73	89.41

EF/LF indicate early/late fusion. The bold indicates the best performance.

followed previous studies [19, 47] in estimating the energy consumption of SNNs by converting FLOPs into energy units. Firstly, the results indicate that the energy consumption of SNNs is significantly lower than that of ANNs. For example, on the CREMA-D dataset, MISNet-XS consumes only 1.32 mJ, achieving 2.8% higher accuracy than the ANN method, while its energy consumption is less than 1% of the ANN. In addition, compared to similar SNN-base like SMMT, MISNet also exhibits a clear energy efficiency advantage. For instance, on the CREMA-D dataset, SMMT consumes 5.78 mJ with an accuracy of 62.50%, while MISNet-XS consumes only 1.32 mJ and achieves an accuracy of 66.58%. MISNet-S, with an energy consumption of 2.41 mJ, further improves accuracy to 68.72%. These results further confirm that MISNet demonstrates excellent energy-efficient when handling multimodal data, thanks to its use of sparse discrete spikes in feature maps for forward propagation.

4.5 Ablation Study

Ablation Study on MLIF. Figure 6 presents the results of the ablation study on the internal structure of MLIF, applying various modification schemes. Specifically, Ab1 denotes the removal of multimodal charging in MLIF, directly performing multimodal fusion during the charging phase, effectively reducing it to a single LIF neuron for processing multimodal inputs; Ab2 indicates the removal of both multimodal fusion and firing operations, resulting in two independent LIF neurons; Ab3 represents the default MLIF neuron.



Ab1 MLIF Neuron w/o Multimodal Charge

Ab2 MLIF Neuron w/o Multimodal Fusion and Multimodal Fire

Ab3 MLIF Neuron

MLIF Internal Results

Case		Fusion	CREMA-D MISNet-L	Ub8k-AV MISNet-S
Ab1	Ab1.1	*Concat*	65.24	93.09
	Ab1.2	*MLP*	67.27	93.42
	Ab1.3	*Attention*	68.91	93.56
Ab2	Ab2.1	*Concat*	62.55	91.21
	Ab2.2	*MLP*	63.81	92.73
	Ab2.3	*Attention*	61.23	89.77
Ab3	Ab3.1	*Concat*	75.22	97.76
	Ab3.2	***MLP***	**78.24**	**98.01**
	Ab3.3	*Attention*	74.21	96.35

Fig. 6. Ablation study of MLIF Internal: Ab1–Ab3 represent the progressive removal or replacement of certain operations within MLIF, followed by observing their impact on the final classification results.

In Ab1.1 through Ab1.3, we explore different fusion methods during the charging phase, including Concat, MLP, and Attention. The results in Figure 6 indicate that the default MLIF (Ab3) achieves the best performance. When switching to Ab1, accuracy decreases as $C[t]_a$ and $C[t]_v$ are directly fused into a single membrane potential for discharge and reset. Although an accuracy of 93.08% is still achieved on the Ub8k-AV dataset, this is 4.2% lower than MLIF; this performance gap persists even with MLP or Hadamard fusion methods. The CREMA-D dataset shows a similar trend, further confirming that spike firing through multimodal membrane potential accumulation is superior to direct fusion of inputs into a single LIF.

The results for Ab2 indicate that removing both multimodal fusion and firing functions leads to two independent LIF neurons firing spikes, with the poorest performance. As expected, the two modalities fire independently, lacking a coordination mechanism, resulting in suboptimal performance. Furthermore, as shown in Figure 6, Ab2.1–Ab2.3, testing any fusion method within Ab3 yields poor results.

In Ab3.1–Ab3.3, we explore the effects of different fusion methods when fusion $U[t]_a$ and $U[t]_v$ into $U[t]$. In addition to the default Concat method, we also test MLP and Attention. The results show that MLP fusion (Ab3.2) achieves the best performance, with consistent results across the CREMA-D and Ub8k-AV datasets. Specifically, using the MISNet-L model, accuracy on the CREMA-D dataset reaches 78.24%, while on the Ub8k-AV dataset, it reaches 98.01%, further validating the effectiveness of different fusion methods for synchronizing multimodal membrane potentials in MLIF.

Ablation Study on Fusion Methods. To demonstrate the effectiveness of multimodal spike interaction in MISNet, we aim to show that its performance is not influenced by the fusion method used in the final classification. The results are shown in Table 6. Therefore, in Equation (13), we replaced the Concat fusion method with Sum, Hadamard, MLP, and Attention and conducted tests. Compared with spike-based SEW-ResNet-18 [8] (which uses two independent encoders to extract

Table 7. Ablation Study of MLIF Neuron in MISNet-L on CREMA-D at $T = 6$, Demonstrating That MLIF Significantly Enhances Effectiveness with Minimal Parameter Increase, Even without the *SEW* Block for Each Stage

Ab.	*SEW* block	MLIF	Param (M)	Acc (%)
#1	-	-	3.11	67.88
#2	✓51	-	19.70	62.55
#3	-	✓51	3.12	72.72
#4	✓51	✓51	19.68	**76.19**

The bold indicates the best performance.

Table 8. Ablation Study of MLIF Neuron and Training Loss in MISNet-L on the CREMA-D at $T = 6$, Demonstrating Significant Performance Improvements with the Addition of MLIF and Further Enhancements When Combined with $\mathcal{L}_a + \mathcal{L}_b$

Ab.	MLIF	$\mathcal{L}_v + \mathcal{L}_a$	Acc (%)
#1	-	-	56.42
#2	-	✓51	62.55
#3	✓51	-	66.80
#4	✓51	✓51	**76.19**

spike features and examines **Early Fusion (EF)** and **Late Fusion (LF)** effects), the experimental results show that MISNet performs well across all tested fusion methods, achieving the highest accuracy of 99.58% with Attention. Additionally, Sum, Hadamard, and Concat achieved accuracies of 98.34%, 99.18%, and 98.09%, respectively, significantly outperforming SEW-ResNet-18. Moreover, MISNet-S shows a notable advantage when using Sum, improving by 8.11% and 12.62% compared to SEW-ResNet18-EF and SEW-ResNet18-LF, respectively. These results indicate that MISNet's efficiency does not rely on a specific fusion method, further validating that our proposed multimodal spike interaction method outperforms independent spike emission approaches.

Ablation Study on Multimodal Spike Interaction. The multimodal spike interaction module, *InteractCell*, consists of two components: the *SEW* block and MLIF. To explore the relative contributions of these components, we conduct ablation experiments within the MISNet-L architecture, removing the *SEW* block and MLIF individually and observing their impact on performance. The results, shown in Table 7, provide a detailed view of the model's performance under different configurations. We find that even with the *SEW* block residuals removed, the model still achieves an accuracy of 72.72% by retaining only the MLIF module. In contrast, removing the MLIF module and retaining only the *SEW* block causes accuracy to drop to 62.55%. This indicates that MLIF contributes more significantly than the *SEW* block in multimodal spike interaction. Additionally, although MISNet-L, as a deeper architecture, should theoretically benefit from the addition of *SEW* block residuals, our results show that even with a deeper structure, removing MLIF leads to a sharp drop in performance, underscoring the critical role of MLIF in multimodal interaction.

Additionally, to further validate the effectiveness of MLIF, we explore its impact in conjunction with the regularization loss during the training of MISNet-L. We conduct experiments by individually removing these two modules. The results are shown in Table 8. When the regularization loss $\mathcal{L}_a + \mathcal{L}_v$ is removed, the model's performance declines but still achieves an accuracy of 66.8%. In contrast, if only the regularization loss is retained while removing MLIF, the performance drops to 62.55%. If both components are removed simultaneously, the model's final performance is just 56.42%. These results further demonstrate the effectiveness of MLIF in enhancing the performance of multimodal SNNs during training.

Ablation Study on Timestep. Additionally, we explore enhancing performance by increasing the number of frames, specifically by increasing T. The evaluation results for T set to 5 and 6 are shown in Table 9, where we can see that performance significantly improves with the increase of T. Specifically, MISNet-S achieved accuracy of 69.99% and 70.12% on the CREMA-D dataset for $T = 5$

Table 9. Ablation Study of Timesteps in MISNet-L on CREMA-D and AVE

Architecture	Timestep	Acc (%)	
		CREMA-D	AVE
MISNet-S	5	69.99	65.22
	6	**70.12**	**66.81**
MISNet-L	5	76.19	67.52
	6	**76.38**	**68.81**
MISNet-XL	5	78.29	68.31
	6	**78.56**	**69.42**

It is demonstrated that increasing key frames can indeed significantly enhance the performance of SNNs on audio-video datasets. The bold indicates the best performance.

Table 10. Ablation Study of MLIF Neuron in MISNet-L on CREMA-D

(β_a, β_v)	Acc (%)	(β_a, β_v)	Acc (%)
(0.1, 0.9)	76.11	(0.6, 0.4)	76.11
(0.2, 0.8)	76.71	(0.7, 0.3)	75.51
(0.3, 0.7)	75.30	(0.8, 0.2)	73.82
(0.4, 0.6)	76.11	(0.9, 0.1)	72.31
(0.5, 0.5)	**76.38**		

Various combinations of (β_a, β_v) were tested during training, with results indicating that the optimal setting is $(\beta_a, \beta_v) = (0.5, 0.5)$. The bold indicates the best performance.

and $T = 6$, respectively. Meanwhile, the performance of MISNet-L also continued to improve with the increase in T, ultimately reaching 76.38%. This trend was similarly observed on the AVE dataset.

Ablation Study on β_a and β_v. During training, the hyper-parameters β_a and β_v have a significant impact on performance. To explore the effect of different combinations of β_a and β_v on performance, we evaluate various combinations and their effects on model performance. The results are shown in Table 10, indicating that different combinations of β_a and β_v have a notable impact on the outcomes. For example, increasing β_a led the model to overly rely on the audio modality during training, resulting in faster convergence but insufficient performance in handling visual information, ultimately affecting the final classification accuracy.

Ablation Study on SEW Blocks. The experimental results in Table 11 demonstrate that MISNet with *SwinT* blocks consistently outperforms its *SEW* counterpart across datasets, including CIFAR10-AV, Ub8k-AV, and CREMA-D. Notably, on Ub8k-AV, MISNet-XL achieved 98.97% accuracy, and on CREMA-D, it reached 77.56%. The average accuracy further highlights the advantage, with MISNet-XL achieving 69.10% compared to SEW's 68.04%. These results underscore the *SwinT* block's effectiveness and MISNet's adaptability to both CNN and ViT-based architectures.

4.6 Qualitative Analysis

t-SNE Visualization. To demonstrate that the interaction strategy enhances the representational capability of spike features, Figure 7 illustrates the spike distributions of SEW-ResNet and MISNet-L after Concat fusion on the CREMA-D dataset, visualized using t-SNE [37]. The visualization displays six categories from CREMA-D, each represented by a distinct color, with "sad" shown in yellow. The visualizations were performed at $t = 1$, $t = 2$, and $t = 3$. We are surprised to find that MISNet can effectively distinguish the six categories using only discrete features. Despite these discrete representations being sparse and discrete, they do not hinder MISNet-L from exhibiting high intra-category compactness and inter-category separability as early as $t = 1$, a trend that continues through $t = 2$ and $t = 3$. In contrast, SEW-ResNet, which also uses spike features, demonstrates poorer performance. This observation indirectly validates our hypothesis that the multimodal interaction fusion strategy based on MLIF can significantly enhance the quality of spike features, thereby improving model performance.

Spike Activity. To further demonstrate the effectiveness of MLIF in multimodal spike emission and in alleviating the imbalance between modalities, we record the spiking activity of the 17th

Table 11. The Ablation Study on the *SEW* Blocks

Block type	Method	Dataset			
		CIFAR10-AV	Ub8k-AV	CREMA-D	AVE
SEW blocks	MISNet-XS	99.53	97.71	66.45	58.44
	MISNet-S	99.55	98.09	68.72	65.31
	MISNet-L	99.68	98.61	75.22	67.24
	MISNet-XL	99.86	98.68	77.14	68.04
SwinT blocks	MISNet-XS	99.63	98.13	67.87	58.80
	MISNet-S	99.65	98.84	69.73	66.42
	MISNet-L	99.78	98.89	76.61	68.92
	MISNet-XL	**99.88**	**98.97**	**77.56**	**69.10**

Best results are in bold, runner-up results are marked in gray. *SwinT* blocks represent Swin Transformer blocks.

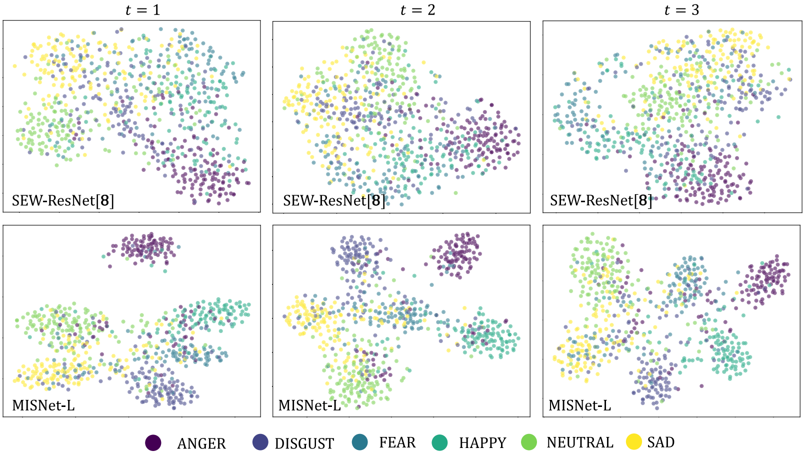


Fig. 7. Qualitative comparison of t-SNE visualization of fused spikes on the CREMA-D test data. *SEW*-ResNet employs LF. Different colors represent different categories. MISNet, using only discrete spike features, achieves clear spatial separation of each category at $t = 1$.

sample in the CREMA-D test set. For comparison, we also provide the spiking activity of MISNet without MLIF. The results in Figure 8 show that MLIF effectively coordinates the spiking activity of both modalities, achieving a relative balance. In contrast, the spiking activity without MLIF exhibits significant imbalance and disorder, with the audio modality firing much more frequently while the video modality fires significantly less.

Spike Features. As shown in Figure 9, we conducted a case study of the spike feature maps on the Ub8k-AV and AVE datasets, comparing them with SMMT. For the Ub8k-AV dataset, we set $T = 1$, while for the AVE dataset, we set $T = 5$ to visualize the spike feature maps in Stage 1. The results of the spike features indicate that, compared to SMMT, which uses independent dual encoders, MISNet adopts an interactive dual-encoder design. This allows the two pathways to work collaboratively, effectively activating spike features in key regions and successfully capturing critical information from input images and audio. In contrast, SMMT, due to the lack of interaction between its encoders, fails to achieve similar performance in these aspects. This further demonstrates that MISNet, through

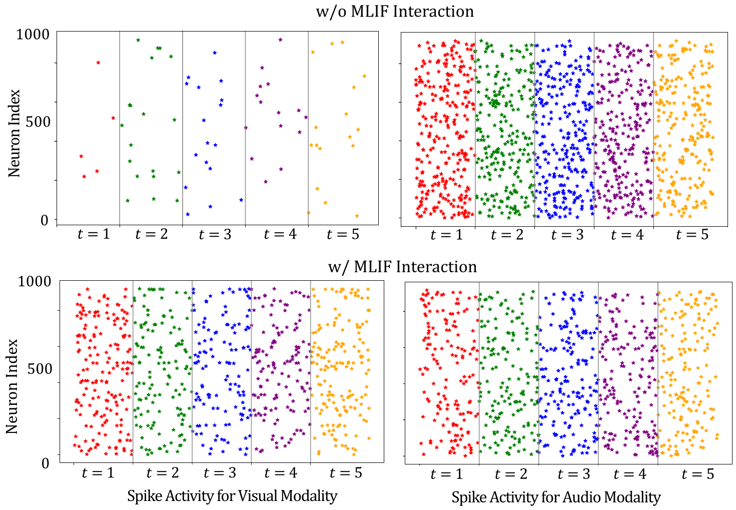


Fig. 8. Qualitative comparison of final-stage spike activity at each timestep t of MISNet with and without MLIF. The vertical axis represents the neuron index. It indicates that, compared to the w/o MLIF interaction, with MLIF interaction can effectively improve the spike firing rate at each t in the visual modality, thereby alleviating the issue of spike firing imbalance.

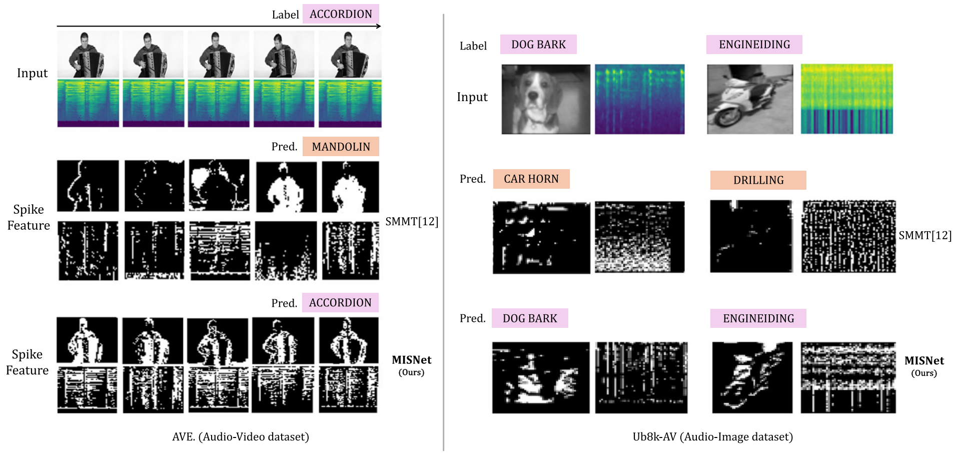


Fig. 9. Case study visualizes the discrete spike feature maps on the Ub8k-AV ($T = 1$) and AVE ($T = 5$) datasets, demonstrating that the spike features of MISNet effectively focus on key regions. Notably, in the more challenging AVE dataset, SMMT [12] fails to effectively focus on these important regions, resulting in suboptimal classification performance. In contrast, MISNet is able to capture the most critical classification areas within the video and accurately identify key audio segments.

its interactive design, can achieve superior performance in AVC tasks by relying solely on discrete spike features, while avoiding the resource overhead associated with floating-point representations in traditional ANNs.

5 Conclusion

In this work, we propose an effective interactive architecture based on SNNs, called MISNet, aimed at efficiently conducting multimodal audiovisual spike learning. MISNet enhances the spike emission effects of multimodal data through multi-round interactions using MLIF during forward propagation. Additionally, to further optimize the effective extraction of spike features for each modality during training, we introduce an independent optimization regularization mechanism. We mainly address two issues: (1) effectively extracting spike features from SNNs in AVC tasks; and (2) addressing the modality imbalance problem caused by the heterogeneous distribution of different modalities and the spike firing characteristics of SNN neurons' membrane potential, which prior works often overlook. Experimental results demonstrate that MISNet performs well on multiple audiovisual datasets; further results on audio-video datasets provide additional evidence of MISNet's effectiveness, and ablation studies validate the efficacy of MLIF. In the future, we will continue to explore the application of SNNs in multimodal learning, aiming to design more general solutions.

References

[1] Malyaban Bal and Abhronil Sengupta. 2024. SpikingBERT: Distilling BERT to train spiking language models using implicit differentiation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, 10998–11006.

[2] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. 2014. Crema-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing* 5, 4 (2014), 377–390.

[3] Yuhao Cheng, Xiaoguang Zhu, Jiuchao Qian, Fei Wen, and Peilin Liu. 2022. Cross-modal graph matching network for image-text retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications* 18, 4 (2022), 1–23.

[4] Chaoteng Duan, Jianhao Ding, Shiyan Chen, Zhaofei Yu, and Tiejun Huang. 2022. Temporal effective batch normalization in spiking neural networks. In *Proceedings of the Advances in Neural Information Processing Systems*, Vol. 35, 34377–34390.

[5] Florian Eyben, Stavros Petridis, Björn Schuller, George Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 2011. Audiovisual classification of vocal outbursts in human conversation using long-short-term memory networks. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 5844–5847.

[6] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. 2023. PMR: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20029–20038.

[7] Wei Fang, Yanqi Chen, Jianhao Ding, Zhaofei Yu, Timothée Masquelier, Ding Chen, Liwei Huang, Huihui Zhou, Guoqi Li, and Yonghong Tian. 2023. Spikingjelly: An open-source machine learning infrastructure platform for spike-based intelligence. *Science Advances* 9, 40 (2023), 1480.

[8] Wei Fang, Zhaofei Yu, Yanqi Chen, Tiejun Huang, Timothée Masquelier, and Yonghong Tian. 2021. Deep residual learning in spiking neural networks. In *Proceedings of the Advances in Neural Information Processing Systems*, Vol. 34, 21056–21069.

[9] Shiming Ge, Fanzhao Lin, Chenyu Li, Daichi Zhang, Weiping Wang, and Dan Zeng. 2022. Deepfake video detection via predictive representation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications* 18, 2 (2022), 1–21.

[10] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. 2024. Mastering deepfake detection: A cutting-edge approach to distinguish GAN and diffusion-model images. *ACM Transactions on Multimedia Computing, Communications, and Applications* 20, 11 (2024), 1551–6857.

[11] Dan Guo, Kun Li, Bin Hu, Yan Zhang, and Meng Wang. *2024. Benchmarking micro-action recognition: Dataset, methods, and applications. IEEE Transactions on Circuits and Systems for Video Technology* 34, 7 (2024), 6238–6252.

[12] Lingyue Guo, Zeyu Gao, Jinye Qu, Suiwu Zheng, Runhao Jiang, Yanfeng Lu, and Hong Qiao. 2024. Transformer-based spiking neural networks for multimodal audio-visual classification. *IEEE Transactions on Cognitive and Developmental Systems* 16, 3 (2024), 1077–1086.

[13] Ruohao Guo, Xianghua Ying, Yaru Chen, Dantong Niu, Guangyao Li, Liao Qu, Yanyu Qi, Jinxing Zhou, Bowei Xing, Wenzhen Yue, et al. 2023. Audio-visual instance segmentation. arXiv:2310.18709. Retrieved from https://arxiv.org/abs/2310.18709

[14] Yufei Guo, Yuanpei Chen, Xiaode Liu, Weihang Peng, Yuhan Zhang, Xuhui Huang, and Zhe Ma. 2024. Ternary spike: Learning ternary spikes for spiking neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, 12244–12252.

[15] Yufei Guo, Yuanpei Chen, Liwen Zhang, Xiaode Liu, Yinglei Wang, Xuhui Huang, and Zhe Ma. 2022. IM-Loss: Information maximization loss for spiking neural networks. In *Proceedings of the Advances in Neural Information Processing Systems*, Vol. 35, 156–166.

[16] YuFei Guo, Yuanpei Chen, Liwen Zhang, YingLei Wang, Xiaode Liu, Xinyi Tong, Yuanyuan Ou, Xuhui Huang, and Zhe Ma. 2023. InfLoR-SNN: Reducing information loss for spiking neural networks. In *Proceedings of the European Conference on Computer Vision*, 36–52.

[17] Yufei Guo, Yuhan Zhang, Yuanpei Chen, Weihang Peng, Xiaode Liu, Liwen Zhang, Xuhui Huang, and Zhe Ma. 2023. Membrane potential batch normalization for spiking neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19420–19430.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–778.

[19] Yulong Huang, Xiaopeng Lin, Hongwei Ren, Yue Zhou, Zunchang Liu, Haotian Fu, Biao Pan, and Bojun Cheng. 2024. CLIF: Complementary leaky integrate-and-fire neuron for spiking neural networks. In *Proceedings of the International Conference on Machine Learning*. Retrieved from https://openreview.net/forum?id=yY6N89IlHa

[20] Yuhang Li, Yufei Guo, Shanghang Zhang, Shikuang Deng, Yongqing Hai, and Shi Gu. 2021. Differentiable spike: Rethinking gradient-descent for training spiking neural networks. In *Proceedings of the Advances in Neural Information Processing Systems, Vol.* 34, 23426–23439.

[21] Zhangbin Li, Dan Guo, Jinxing Zhou, Jing Zhang, and Meng Wang. 2024. Object-aware adaptive-positivity learning for audio-visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3306–3314.

[22] Zhangbin Li, Jinxing Zhou, Jing Zhang, Shengeng Tang, Kun Li, and Dan Guo. 2024. Patch-level sounding object tracking for audio-visual question answering. arXiv:2412.10749. Retrieved from https://arxiv.org/abs/2412.10749

[23] Qianhui Liu, Dong Xing, Lang Feng, Huajin Tang, and Gang Pan. 2022. Event-based multimodal spiking neural network with attention mechanism. In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*, 8922–8926.

[24] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Zechao Li, Qi Tian, and Qingming Huang. 2023. Entity-enhanced adaptive reconstruction network for weakly supervised referring expression grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 3 (2023), 3003–3018.

[25] Wolfgang Maass. 1997. Networks of spiking neurons: The third generation of neural network models. *Neural Networks* 10, 9 (1997), 1659–1671.

[26] Yuxin Mao, Xuyang Shen, Jing Zhang, Zhen Qin, Jinxing Zhou, Mochu Xiang, Yiran Zhong, and Yuchao Dai. 2024. TAVGBench: Benchmarking text to audible-video generation. arXiv:2404.14381. Retrieved from https://arxiv.org/abs/2404.14381

[27] Adnan Mehonic and Anthony J. Kenyon. 2022. Brain-inspired computing needs a master plan. *Nature* 604, 7905 (2022), 255–260.

[28] Qingyan Meng, Mingqing Xiao, Shen Yan, Yisen Wang, Zhouchen Lin, and Zhi-Quan Luo. 2022. Training high-performance low-latency spiking neural networks by differentiation on spike representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12444–12453.

[29] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention bottlenecks for multimodal fusion. In *Proceedings of the Advances in Neural Information Processing Systems*, Vol. 34, 14200–14213.

[30] Shanbao Qiao, Neal N. Xiong, Yongbin Gao, Zhijun Fang, Wenjun Yu, Juan Zhang, and Xiaoyan Jiang. 2023. Self-supervised learning of depth and ego-motion for 3D perception in human computer interaction. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 2 (2023), 1–21.

[31] Kaushik Roy, Akhilesh Jaiswal, and Priyadarshini Panda. 2019. Towards spike-based machine intelligence with neuromorphic computing. *Nature* 575, 7784 (2019), 607–617.

[32] Xuyang Shen, Dong Li, Jinxing Zhou, Zhen Qin, Bowen He, Xiaodong Han, Aixuan Li, Yuchao Dai, Lingpeng Kong, Meng Wang, et al. 2023. Fine-grained audible video description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10585–10596.

[33] Xinyu Shi, Zecheng Hao, and Zhaofei Yu. 2024. SpikingResformer: Bridging ResNet and vision transformer in spiking neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5610–5619.

[34] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. 2018. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision*, 247–263.

[35] Yao-Hung Hubert Tsai, Paul PuLiang, AmirZadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations. In *Proceedings of the International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=rygqqsA9KX

[36] Yunbin Tu, Liang Li, Li Su, Zheng-Jun Zha, and Qingming Huang. 2024. SMART: Syntax-calibrated multi-aspect relation transformer for change captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 7 (2024), 4926–4943.

[37] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 11 (2008), 2579–2605.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems*, Vol. 30.

[39] Hao Wang, Zheng-Jun Zha, Liang Li, Xuejin Chen, and Jiebo Luo. 2023. Semantic and relation modulation for audio-visual event localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 6 (2023), 7711–7725.

[40] Jun Wu, Tianliang Zhu, Jiahui Zhu, Tianyi Li, and Chunzhi Wang. 2023. A optimized BERT for multimodal sentiment analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications* 19, 2 (2023), 1–12. DOI: https://doi.org/10.1145/3566126

[41] Yujie Wu, Lei Deng, Guoqi Li, Jun Zhu, Yuan Xie, and Luping Shi. 2019. Direct training for spiking neural networks: Faster, larger, better. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, 1311–1318.

[42] Xiongye Xiao, Gengshuo Liu, Gaurav Gupta, Defu Cao, Shixuan Li, Yaxing Li, Tianqing Fang, Mingxi Cheng, and Paul Bogdan. 2024. Neuro-inspired information-theoretic hierarchical perception for multimodal learning. In *Proceedings of the International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=Z9AZsU1Tju

[43] Xun Yang, Tianyu Chang, Tianzhu Zhang, Shanshan Wang, Richang Hong, and Meng Wang. 2024. Learning hierarchical visual transformation for domain generalizable visual matching and recognition. *International Journal of Computer Vision* 132 (2024), 1–27.

[44] Xun Yang, Shanshan Wang, Jian Dong, Jianfeng Dong, Meng Wang, and Tat-Seng Chua. 2022. Video moment retrieval with cross-modal neural architecture search. *IEEE Transactions on Image Processing* 31 (2022), 1204–1216.

[45] Xun Yang, Jianming Zeng, Dan Guo, Shanshan Wang, Jianfeng Dong, and Meng Wang. 2024. Robust video question answering via contrastive cross-modality representation learning. *Science China Information Sciences* 67 (2024), 1–16.

[46] Man Yao, Jiakui Hu, Zhaokun Zhou, Li Yuan, Yonghong Tian, Bo Xu, and Guoqi Li. 2023. Spike-driven transformer. In *Proceedings of the Advances in Neural Information Processing Systems*, Vol. 36, 64043–64058.

[47] Man Yao, Guangshe Zhao, Hengyu Zhang, Yifan Hu, Lei Deng, Yonghong Tian, Bo Xu, and Guoqi Li. 2023. Attention spiking neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 8 (2023), 9393–9410.

[48] Ying Zeng, Sijie Mai, Wenjun Yan, and Haifeng Hu. 2024. Multimodal reaction: Information modulation for cross-modal representation learning. *IEEE Transactions on Multimedia* 26 (2024), 2178–2191.

[49] Beichen Zhang, Liang Li, Shuhui Wang, Shaofei Cai, Zheng-Jun Zha, Qi Tian, and Qingming Huang. 2024. Inductive state-relabeling adversarial active learning with heuristic clique rescaling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 12 (2024), 9780–9796.

[50] Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. 2024. Multimodal representation learning by alternating unimodal adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27456–27466.

[51] Youhui Zhang, Peng Qu, Yu Ji, Weihao Zhang, Guangrong Gao, Guanrui Wang, Sen Song, Guoqi Li, Wenguang Chen, Weimin Zheng, et al. 2020. A system hierarchy for brain-inspired computing. *Nature* 586, 7829 (2020), 378–384.

[52] Zhihuang Zhang, Jintao Zhao, Changyao Huang, and Liang Li. 2022. Learning visual semantic map-matching for loosely multi-sensor fusion localization of autonomous vehicles. *IEEE Transactions on Intelligent Vehicles* 8, 1 (2022), 358–367.

[53] Pengcheng Zhao, Jinxing Zhou, Dan Guo, Yang Zhao, and Yanxiang Chen. 2024. Multimodal class-aware semantic enhancement network for audio-visual video parsing. arXiv:2412.11248. Retrieved from https://arxiv.org/abs/2412.11248

[54] Jinxing Zhou, Dan Guo, Ruohao Guo, Yuxin Mao, Jingjing Hu, Yiran Zhong, Xiaojun Chang, and Meng Wang. 2024. Towards open-vocabulary audio-visual event localization. arXiv:2411.11278. Retrieved from https://arxiv.org/abs/2411.11278

[55] Jinxing Zhou, Dan Guo, Yuxin Mao, Yiran Zhong, Xiaojun Chang, and Meng Wang. 2024. Label-anticipated event disentanglement for audio-visual video parsing. In *Proceedings of the European Conference on Computer Vision*, 1–22.

[56] Jinxing Zhou, Dan Guo, and Meng Wang. 2023. Contrastive positive sample propagation along the audio-visual event line. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45 (2023), 7239–7257.

[57] Jinxing Zhou, Dan Guo, Yiran Zhong, and Meng Wang. 2024. Advancing weakly-supervised audio-visual video parsing via segment-wise pseudo labeling. *International Journal of Computer Vision* 132 (2024), 1–22.

[58] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, et al. 2023. Audio-visual segmentation with semantics. arXiv:2301.13190. Retrieved from https://arxiv.org/abs/2301.13190

[59] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. 2022. Audio-visual segmentation. In *Proceedings of the European Conference on Computer Vision*, 386–403.

[60] Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. 2021. Positive sample propagation along the audio-visual event line. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8436–8444.

[61] Ziheng Zhou, Jinxing Zhou, Wei Qian, Shengeng Tang, Xiaojun Chang, and Dan Guo. 2024. Dense audio-visual event localization under cross-modal consistency and multi-temporal granularity collaboration. arXiv:2412.12628. Retrieved from https://arxiv.org/abs/2412.12628

[62] Zhaokun Zhou, Yuesheng Zhu, Chao He, Yaowei Wang, Shuicheng Yans, Yonghong Tian, and Li Yuan. 2023. Spikformer: When spiking neural network meets transformer. In *Proceedings of the International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=frE4fUwz_h

[63] Yaoyu Zhu, Wei Fang, Xiaodong Xie, Tiejun Huang, and Zhaofei Yu. 2023. Exploring loss functions for time-based training strategy in spiking neural networks. In *Proceedings of the Advances in Neural Information Processing Systems, Vol*. 36, 65366–65379.

Received 7 November 2024; revised 15 January 2025; accepted 23 February 2025